NPRDC TR 83-15

APRIL 1983

INFLUENCE OF FALLIBLE ITEM PARAMETERS ON TEST INFORMATION DURING ADAPTIVE TESTING

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN DIEGO, CALIFORNIA

NAVY PERSONNEL RESEARCH
AND
DEVELOPMENT CENTER
San Diego, California 92152

NPRDC TR 83-15

April 1983

INFLUENCE OF FALLIBLE ITEM PARAMETERS ON TEST INFORMATION DURING ADAPTIVE TESTING

C. Douglas Wetzel
James R. McBride

Reviewed by
Martin F. Wiskoff

Released by
James F. Kelly, Jr.
Commanding Officer

Navy Personnel Research and Development Center
San Diego, California, 92152

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
NPRDC TR 83-15

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
INFLUENCE OF FALLIBLE ITEM PARAMETERS ON TEST INFORMATION DURING ADAPTIVE TESTING

5. TYPE OF REPORT & PERIOD COVERED
Technical Report
Jun 1981-Sep 1982

6. PERFORMING ORG. REPORT NUMBER
12-82-7

7. AUTHOR(s)

C. Douglas Wetzel
James R. McBride

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Navy Personnel Research and Development Center
San Diego, California 92152

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
PE 62763N
CF63-521-080-101-04.12

11. CONTROLLING OFFICE NAME AND ADDRESS

Navy Personnel Research and Development Center
San Diego, California 92152

12. REPORT DATE
April 1983

13. NUMBER OF PAGES
26

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)

UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Computerized adaptive testing
Tailored testing
Psychological testing
Personnel testing
Computer simulation
STRADAPTIVE test
Maximum information test
Item response theory

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Computer simulation was used to assess the effects of item parameter estimation errors on different item selection strategies used in adaptive and conventional testing. To determine whether these effects reduced the advantages of certain optimal item selection strategies, simulations were repeated in the presence and absence of item parameter estimation errors. Results showed that item parameter estimation errors had little effect on the efficiency and measurement precision of the adaptive test item selection strategies studied. Strategies that explicitly made optimal use of item parameters for item selection were superior to a less optimal strategy, even when item parameters were fallibly estimated. It appears that errors in the item parameter estimates do not reduce the psychometric advantages of these "optimal" strategies. Item selection strategies that explicitly employ optimization criteria should be regarded as preferable to simpler strategies that do not. Further development of psychometric procedures for the CAT system should focus on the former type of strategy.

FOREWORD

A joint-service coordinated effort is in progress to develop a computerized adaptive testing (CAT) system and to evaluate its potential for use in the military entrance processing stations as a replacement for the Armed Services Vocational Aptitude Battery (ASVAB) printed tests. The Navy Personnel Research and Development Center has been designated lead laboratory for this effort.

This report describes the role of fallible item parameter estimation techniques on several item selection strategies used during testing. This research was conducted as part of project CF63-521-080-101-04.12 (USMC Computerized Adaptive Testing). It was directed toward technical, professional, and contractor personnel involved in implementing CAT. Previous reports described CAT system functional requirements and schedules and preliminary design considerations (NPRDC Tech. Note 82-22 and Tech. Rep. 82-52).

JAMES F. KELLY, JR.
Commanding Officer

JAMES W. TWEEDDALE
Technical Director

SUMMARY

Problem

The Navy Personnel Research and Development Center is developing a computerized adaptive testing (CAT) system as a possible replacement for the Armed Services Vocational Aptitude Battery. An essential feature of CAT is the tailoring of aptitude tests to the individual, under computer control. This tailoring is accomplished by selecting test items whose psychometric characteristics closely match the apparent ability level of the examinee. A variety of methods to accomplish this matching exists. An important aspect of CAT system development is the choice of the item selection procedure to be used.

Objective

The present research was designed to compare several prominent adaptive testing strategies for item selection in terms of measurement precision and efficiency. Secondary purposes were (1) to assess the effect of item parameter estimation errors on the different item selection strategies, and (2) to determine whether these effects reduced the advantages of the optimal item selection strategies.

Approach

Computer simulation was used to evaluate typical levels of error in the estimation of test item parameters and to assess the measurement precision and efficiency of several item selection strategies, under each of two conditions--the presence and absence of item parameter estimation errors. Four strategies--three adaptive and one conventional--were compared both within and across these two conditions.

Findings

Item parameter estimation errors had little effect on the efficiency and measurement precision of the adaptive test item selection strategies studied. Strategies that made explicit optimal use of item parameters for item selection were superior to a less optimal strategy, even when item parameters were fallibly estimated.

Conclusions

Of the adaptive test item selection strategies compared in this investigation, the two strategies that use test item parametric values to select test items in an optimal fashion were superior to the other strategies studied. It appears that errors in the item parameter estimates do not reduce the psychometric advantages of these "optimal" strategies.

Recommendations

Item selection strategies that explicitly employ optimization criteria should be regarded as preferable to simpler strategies that do not. Further development of psychometric procedures for the CAT system should focus on the former type of strategy.

CONTENTS

Page

INTRODUCTION 1

Problem and Background 1
Item Response Models 1
Adaptive Testing Strategies 3
Classes of Adaptive Testing Strategies 4
Purpose 5

APPROACH 5

Generating Fallible Item Parameter Estimates 5
Simulation of Testing Strategies 6
Dependent Variables. 7

RESULTS 7

Item Pool Characteristics 7
Precision of Item Selection. 10
Effective Test Length 13

DISCUSSION 15

CONCLUSIONS. 16

RECOMMENDATIONS. 16

REFERENCES 17

DISTRIBUTION LIST 19

LIST OF FIGURES

1. Item response and item information functions for three sample items 2

2. Scatter plots of true and estimated parameters for a simulated 223-item adaptive testing item bank 9

3. Best possible test information for a 15-item test, calculated from true and estimated item parameters 10

4. Average test information obtained with four test strategies. 11

5. Average test information for each test using true vs. estimated item parameters for item selection 12

6. Adaptive efficiency for each test using true vs. estimated item parameters for item selection 14

7. Relative efficiency and effective test length of three adaptive tests 15

INTRODUCTION

Problem and Background

The Department of Defense is currently developing computerized adaptive testing (CAT) as a potential replacement for the Armed Services Vocational Aptitude Battery (ASVAB) of paper-and-pencil tests used for enlisted personnel selection and classification. CAT entails automated test administration in which the sequence of test items is tailored to each examinee, contingent on his/her responses to earlier items in the sequence. Sequential choice of test items is based on the examinee's performance and on the parameters of a unique item response model previously fitted to each test item. The precision and efficiency of an adaptive test presume that item response model parameters have been accurately estimated.

The essence of adaptive testing is that items are chosen to match item difficulty (and other parameters) to the apparent ability of the examinee. A variety of strategies to accomplish this exists. Some of these strategies use mathematical optimization techniques for sequential choice of test items, while others use simple but suboptimal branching rules. It is generally accepted that the optimization-based strategies are superior to the simpler ones in measurement precision and efficiency. This is certainly the case when the item parameters contain no error. This result has been obtained by means of both analytic studies and Monte Carlo simulation (cf. Crichton, 1981).

Item Response Models

In traditional test theory (e.g., Gulliksen, 1950), the psychometric characteristics of different test items were expressed as "item difficulty" (the proportion of the norm group who answered the item correctly) and "discriminating power" (the correlation between total test score and performance (correct or incorrect) on the item).

Traditional test theory is not particularly useful in the context of adaptive testing. Item response theory (IRT) (Lord, 1980) is a more modern and useful formulation. In IRT, the psychometric characteristics of different test items are expressed as parameters of mathematical models called item response functions, which express the probability of a correct answer as a function of examinee ability. Figure 1a illustrates three item response functions, each having the same general mathematical form but different parameters. The general form of these models is a three-parameter logistic function:

$$P(t) = c + \frac{(1 - c)}{1 + \exp^{-1.7\, a\, (t - b)}} .$$

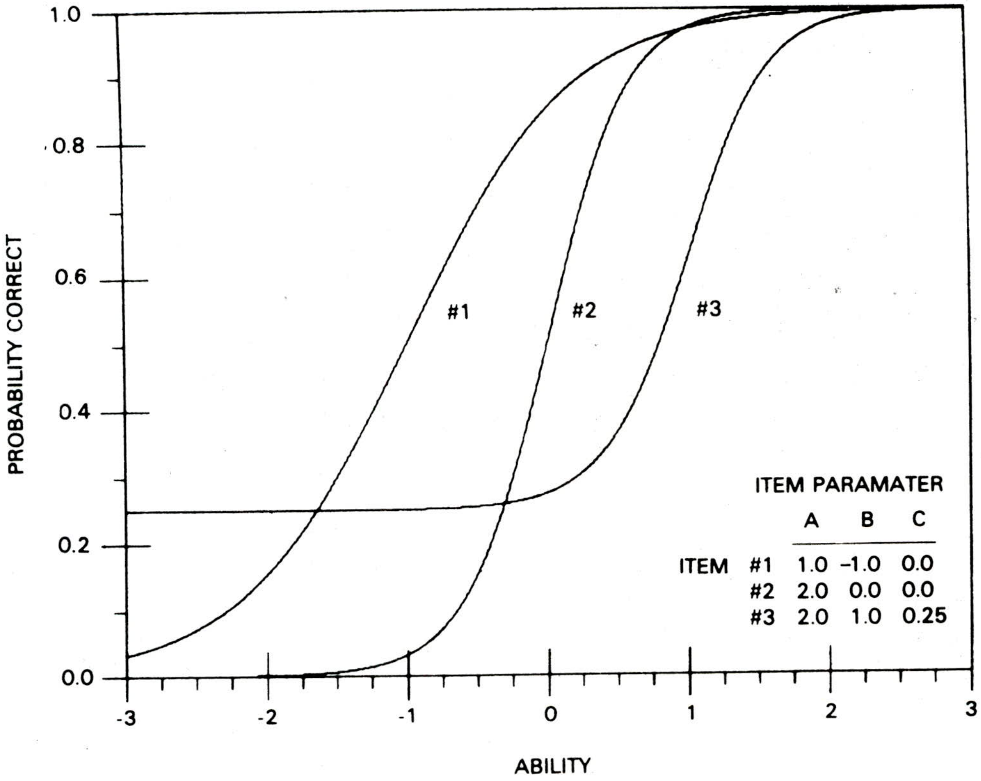
Each item's response function is distinguished from those of other items by a specific set of parameters $\{\underline{a}, \underline{b}, \underline{c}\}$. The equation is interpreted in the following manner:

P(t) is the probability of a correct answer for a given value of t.

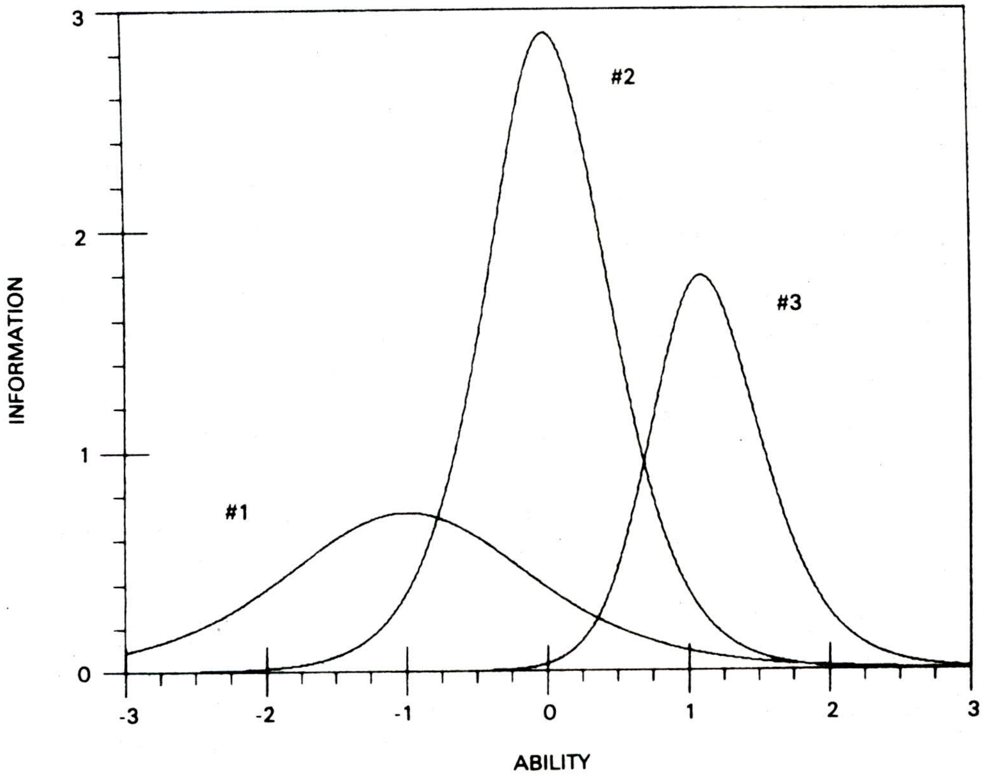
t or theta is a real number that expresses the ability level of the examinee.

$\underline{b}$ is the "difficulty," or threshold, of the item; that is, the ability level at which there is a 0.5 probability of <u>knowing</u> the correct answer.

$\underline{a}$ is the "discrimination" parameter, which is proportional to the maximum slope of the response function.



a. Item response functions.



b. Item information functions.

Figure 1. Item response and item information functions for three sample items.

c — is the lower asymptote parameter: the probability that an examinee of very low ability will answer the item correctly.

The three-item response functions depicted in Figure 1a have different a, b, and c parameters. Item 1 is a relatively "easy" item (b = -1.0), is somewhat discriminating (a = 1.0), and is not susceptible to guessing (c = 0.0). Item 2 is of median difficulty (b = 0.0), is highly discriminating (a = 2.0), and is no more susceptible to guessing than is item 1 (c = 0.0). Item 3 is a difficult, highly discriminating item that is also somewhat susceptible to guessing (b = 1.0, a = 2.0, c = 0.25).

In practical applications of item response theory, item response data are analyzed to estimate the values of response model parameters for each test item. These parameter estimates are subsequently used as a basis for test design (selecting a set of items to form a test for some specific purpose) and for sophisticated ability estimation and test analysis techniques.

Adaptive Testing Strategies

The distinguishing feature of adaptive testing (as opposed to conventional testing) is the tailored selection of test items whose parameters are well matched to the ability level of each individual examinee. This selection is usually based on an estimate of the examinee's ability level (t or theta) and on some function of the item parameters $\{a, b, c\}$. For example, an item might be selected to minimize the value of $|\ t - b\ |$, the absolute difference between ability and the item's difficulty parameter. A more sophisticated rationale is to select the item that maximizes item information (I), which is defined by Birnbaum (1968, p. 449) as

$$I(t) = \frac{\left[\frac{d\,P(t)}{d\,t}\right]^2}{P(t)\,(1 - P(t))} \quad .$$

Item information (I) can also be computed as follows:

$$I = (1.7a)^2 \cdot \frac{\left[e^{1.7a(b-t)}\right] - \left[c\,e^{1.7a(b-t)}\right]}{\left[1 + c\,e^{1.7a(b-t)}\right] \cdot \left[1 + e^{1.7a(b-t)}\right]^2} ,$$

where a, b, and c are the 3-parameter logistic model item parameters and t is examinee ability. Notice that the item information function not only considers the difference (t - b) between ability and item difficulty but also the slope of the item response function and the item's apparent susceptibility to guessing (c).

Figure 1b illustrates the item information functions corresponding to the three item response functions in Figure 1a. By comparing the two, it can be seen that (1) item information is highest in the region of ability where the item response function's slope is steepest, (2) the total area under the information function increases as the a parameter increases (item #1 vs. #2), and (3) the total area under the information function decreases as the c parameter increases (item #2 vs. #3). It should also be noted that the maximum value of item information occurs at an ability level close to each item's b parameter. These features of item information provide a sophisticated and convenient means by which items can be selected during adaptive testing. Such a strategy has been included for evaluation in the present investigation.

The use of a specific rationale for sequential item selection is an important component of any strategy for adaptive testing. Two other components may also be present. One is the method, if any, used to estimate examinee ability during the course of the adaptive test. The other is the test termination criterion, or stopping rule: the rule used to end the test. These components are described below.

1. Estimating Ability. In an adaptive test, the choice of test items is individually tailored to the examinee's performance. In the optimization-based strategies for adaptive testing, examinee performance is characterized by a statistically derived estimate of examinee ability, which is updated during the test. For the purposes of this report, focus is directed to one specific approach to ability estimation, the Bayesian sequential updating technique given by Owen (1969, 1975) and described elsewhere (Jensema, 1977; Urry & Dorans, in press).

2. Test Termination. A conventional test usually stops when the examinee has answered all the questions or when a time limit is reached. The use of a time limit is generally not appropriate for adaptive testing. Instead, an adaptive test might terminate when the examinee has answered a certain number of test questions. An alternative is to stop testing when the examinee's ability estimate has reached a prespecified degree of precision; however, using this criterion results in tests of different length for different examinees. In this report, a fixed-length termination criterion will be employed; variable-length adaptive tests will be examined in a subsequent report.

Classes of Adaptive Testing Strategies

Strategies for adaptive testing may differ in any of the three characteristics discussed above: item selection rationale, ability estimation procedure, and test termination criterion. Consequently, there is a very large number of strategies possible, so large a number it is impractical to conduct an exhaustive evaluation of the alternatives prior to development of the CAT system. Some comparison of strategies is necessary, however. The approach adopted here is to identify classes of adaptive strategies and to limit comparisons to promising representative strategies within each class.

Classification schemes for adaptive testing strategies have been proposed earlier by Weiss (1974), Waters (1974), and McBride (1976a). In the present report, only two broad classes were considered: strategies that employ mathematical optimization criteria for item selection and strategies that do not. An example of an optimization-based strategy is Owen's (1969, 1975) Bayesian sequential tailored testing procedure; an example of the other class of "mechanical" strategies is the stratified adaptive (STRADAPTIVE) procedure proposed by Weiss (1973).

The Owen strategy selects each item by minimizing a function of (1) a Bayesian prior distribution of ability and (2) the item response model parameters of each test item not yet administered. The prior ability distribution is updated after each item is answered.

In actual testing practice, item response model parameters are not known; instead, they are estimated and the estimates contain errors. The effects of these errors on adaptive tests has not been studied adequately. However, several possibilities exist, one of which is that the optimization-based adaptive testing strategies may not be robust in the presence of item parameter estimation errors. A worst-case possibility is that the theoretical measurement advantages of adaptive tests may not hold in the presence of these errors.

Purpose

The purpose of the present research was to compare several prominent adaptive testing strategies to each other and to a conventional test, in terms of measurement precision and efficiency. Secondary purposes were to (1) assess the effects of item parameter estimation errors on different test strategies and (2) determine whether these effects reduce the effectiveness of optimizing item selection strategies. Adaptive test strategies designed to make optimal use of item parameters as a basis for sequential selection of items may be degraded to the extent that item parameters have been fallibly estimated.

APPROACH

A two-stage computer simulation was used to investigate the effects of item parameter estimation error on the psychometric characteristics (e.g., measurement efficiency and precision) of several adaptive and conventional testing strategies. The first-stage simulation produced item parameter estimates, with typical error characteristics, for a simulated item bank constructed in a manner similar to real adaptive test item banks. This item bank was used in the second stage, in which administrations of adaptive and conventional tests were simulated. The psychometric characteristics of each simulated test were recorded and analyzed to assess the effect of item parameter estimation error on the different testing strategies.

Generating Fallible Item Parameter Estimates

A test-item bank was created to approximate an unselected set of items by generating an initial set of 400 "true" item-parameters. Each simulated item was characterized completely by its unique configuration of $\underline{a}$, $\underline{b}$, and $\underline{c}$ parameters, each being generated independently from a different random, uniform rectangular distribution. The $\underline{a}$ parameters (discrimination) were limited to the range 0.2 to 2.0; the $\underline{b}$ parameters (difficulty), to the range -3.0 to +3.0; and the $\underline{c}$ parameters (guessing), to the range 0.0 to 0.3. These "items" were then "administered" to an initial calibration sample of 2000 simulated examinees sampled from a standard normal distribution. The three-parameter logistic model (Birnbaum, 1968) was used to generate simulated binary responses to the test items, using a probability sampling technique often employed for this purpose (e.g., Vale & Weiss, 1975). If a random number drawn from a uniform distribution on the interval (0, 1) was less than the three-parameter logistic model probability of a correct response P(t), then the examinee was credited with a correct answer; otherwise, an incorrect response was specified for the item. This resulted in a matrix of 2000 (persons) by 400 (items) of simulated correct and incorrect item responses.

The responses of the initial calibration sample were analyzed, using the item-parameter estimation program OGIVIA (Urry, 1977). Because of limitations inherent in the computer program, the parameters of the 400 items were estimated for independent subsets of from 50-57 items at a time. This program produced initial estimates of the true item parameters generated earlier. Two separate runs of OGIVIA were used. The first OGIVIA analysis provided a basis for initial culling of the item bank. An item was omitted from the bank if its estimated $\underline{a}$ parameter was less than 0.8 or if the program indicated that the item failed to fit the three-parameter logistic model. A second run of OGIVIA analyzed a new sample of computer-generated responses to the items retained. These responses were obtained from a new sample of 2000 examinees. Following the second run, items were selected for use in the adaptive test simulations. Again, items for which OGIVIA did not return estimates and items with an estimated $\underline{a}$ parameter less than

0.8 were omitted. Also, items with an estimated a parameter greater than 2.3 or an estimated c parameter greater than 0.3 were omitted.

The 186-item bank resulting from this second culling was then analyzed in terms of the test information of optimal 15-item tests constructed at 19 levels of ability. Inspection revealed an irregularly shaped curve, which indicated that the number of items in the bank was inadequate. Accordingly, a second set of 400 items was generated and subjected to the two-stage response generation/item parameter estimation process described above. From this second set, 37 items were selected to supplement the original 186 items, resulting in a final item bank of 223 items. These 37 additional items were selected to make maximum test information more uniform.

Simulation of Testing Strategies

The bank of 223 simulated test items was then used as a basis for computer simulations of conventional and adaptive tests, each with a fixed test length of 15 items. For each test, 1900 examinees were simulated, 100 at each of 19 ability levels between -2.25 and +2.25.

Each testing strategy was simulated twice, once under each of the following item parameter conditions:

1. An "ideal" condition in which the true item parameters were used as a basis for both item selection and test scoring (ability estimation).

2. A "realistic" condition in which the fallible item parameter estimates were used for both item selection and scoring.

In both conditions, the simulation of examinees' responses was based on the simulated true examinee ability and true item parameters. Item responses were scored as correct if a random number drawn from a uniform distribution on the interval (0, 1) was less than the computed probability of a correct response.

Four test strategies were simulated: three adaptive and one conventional. The adaptive test strategies were Owen's Bayesian sequential procedure, Weiss' STRADAPTIVE procedure, and a "hybrid" Bayesian strategy. The conventional test employed a peaked design--a homogeneous distribution of item difficulty. The adaptive tests began with an initial ability estimate of 0.0; the two Bayesian tests assumed an initial prior distribution of ability with mean 0 and variance 1. Testing strategies are described in more detail below.

1. Owen's Bayesian Sequential Test. This adaptive test strategy, proposed by Owen (1969), chooses the item that minimizes the expected value of the posterior variance of the ability distribution. After each item, the ability estimate (distribution) is updated, and the parameters of the Bayes posterior distribution are employed as the parameters of the prior distribution for the next item. In simulating this strategy, both ability estimation and item selection were based on the same set of either true or estimated parameters.

2. Hybrid Bayesian Test. This adaptive test selects the item with approximately the greatest item information at the current ability estimate from one of 36 prearranged information tables. Each table contains a list of test items arranged in descending order of the values of their information functions at the mid-points of a series of narrow intervals (e.g., .125 wide) of ability. In this study, the 36 tables spanned the ability range

from -2.25 to +2.25. The ability estimate was updated after each item, using the same Bayesian ability estimation procedure employed in the Owen's test described above. Thus, test strategy can be considered a "hybrid" between a Bayesian test and a maximum-likelihood, information table look-up method (cf. Lord, 1977; Sympson, Weiss, & Ree, 1982). As with the Owen strategy, both ability estimation and item selection were based on the same set of either true or estimated item parameters.

3. STRADAPTIVE Test. This mechanical adaptive test strategy was originally proposed by Weiss (1973). In a STRADAPTIVE test, the item pool is sorted into "strata" based on the item b parameters and the items within each stratum are arranged in descending order of the values of their a parameter. In this study, there were nine strata, each 0.5 ability units wide, over the range -2.25 to +2.25. Items were selected from the top of the stack in each stratum. After each item, branching to another stratum occurred--branching up one stratum after a correct response and down one stratum after an incorrect response.

4. Peaked Conventional Test. In this study, the conventional test was designed by selecting the 15 most informative items at the central ability value of 0.0. All simulated examinees were administered this same set of 15 items.

Dependent Variables

The dependent variable for the preceding computer simulations was test information, calculated for the test items administered to each examinee. Test information is an index of precision, or of how well a set of items discriminates an ability level from nearby ability levels; the reciprocal of the square root of information is an index of measurement error. It was used here as a measure of the appropriateness of the set of items administered for a given ability level. "Test information" is the sum of the individual "item information" values (see above) for the items administered in a test. These test information values were then averaged over the 100 examinees at each of the 19 levels of true ability. For simulation runs using only true parameters, test information was calculated with the true parameters. For simulation runs using estimated parameters during item-selection and ability estimation, test information was calculated twice--once with true parameter values and once with the estimated parameters.

RESULTS

Item Pool Characteristics

The item-selection procedure resulted in an item pool with characteristics as described in Table 1. Figure 2 shows scatter plots depicting the relation between the true and estimated item parameters, along with the line of perfect correspondence. All three parameters were systematically overestimated, with the best estimates being made for b, the difficulty parameter, and the worst for c, the guessing parameter. This pattern of results is consistent with previous research using Urry's item-parameter estimation programs (Ree, 1979; Schmidt & Gugel, 1976; Swaminathan & Gifford, 1980).

Figure 3 illustrates the effect of overestimation of item parameters on estimated test information. The figure depicts test information for the 15 most informative items at each of 19 levels of ability. It shows that, when both item selection and information calculations were based on estimated parameters, test information was materially overestimated.

Table 1

Comparison of True and Estimated Parameters
For a Selected Pool of 223 Test Items

Characteristic	Item Parameter		
	a	*b*	*c*
Mean:			
True	1.275	0.052	0.135
Estimated	1.527	0.216	0.175
Standard deviation:			
True	0.372	1.262	0.082
Estimated	0.403	1.289	0.066
Bias (estimated minus true)	0.251	0.164	0.039
Root mean square error (true and estimated)	0.325	0.214	0.072
Correlation (true and estimated)	0.861	0.994	0.689
Sq. correlation (true and estimated)	0.742	0.988	0.475

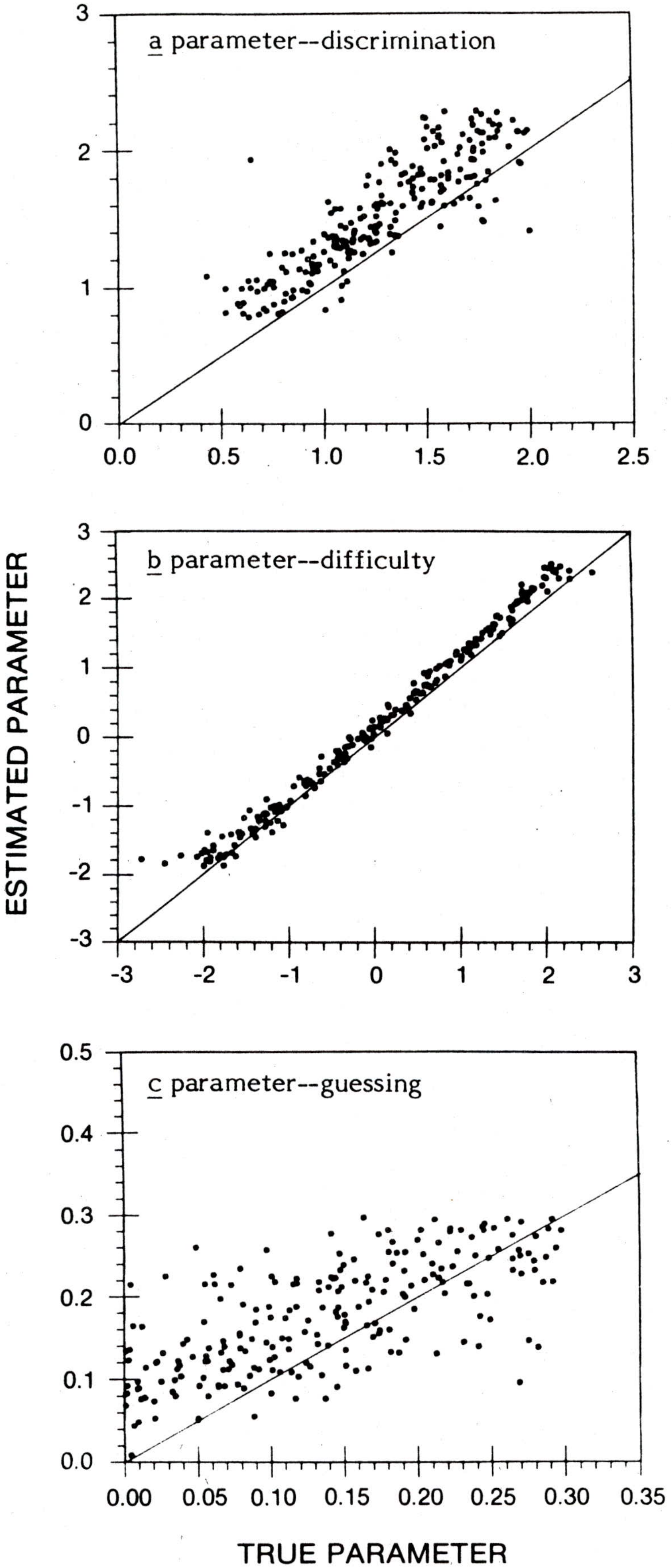


Figure 2. Scatter plots of true and estimated parameters for a simulated 223-item adaptive testing item bank.

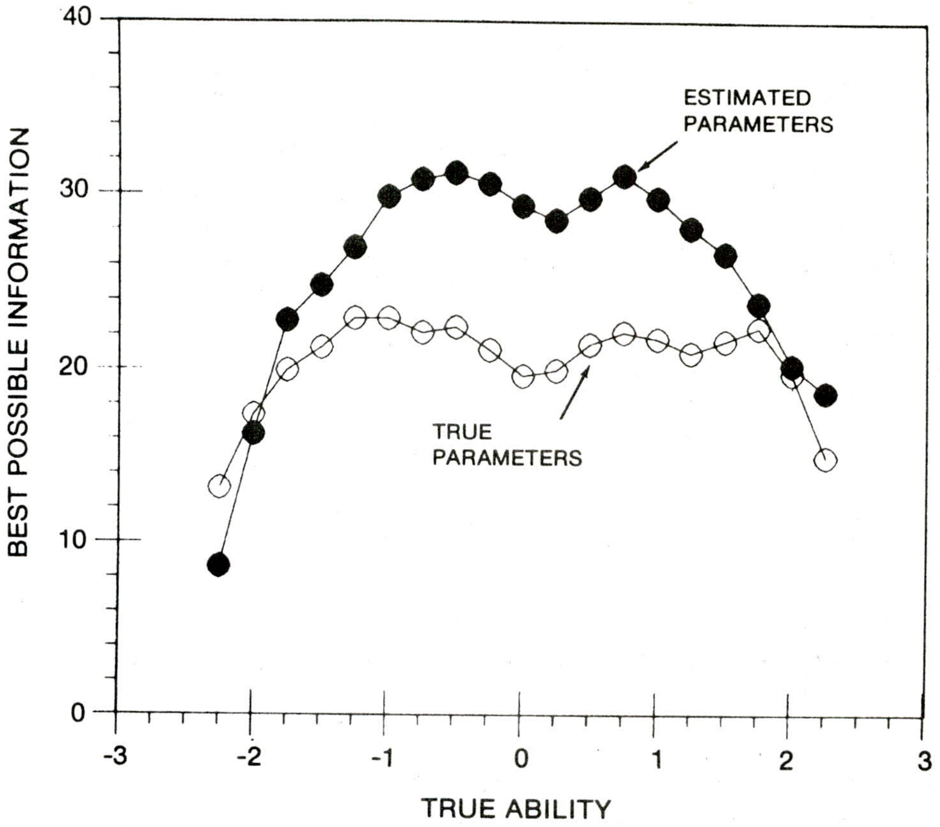
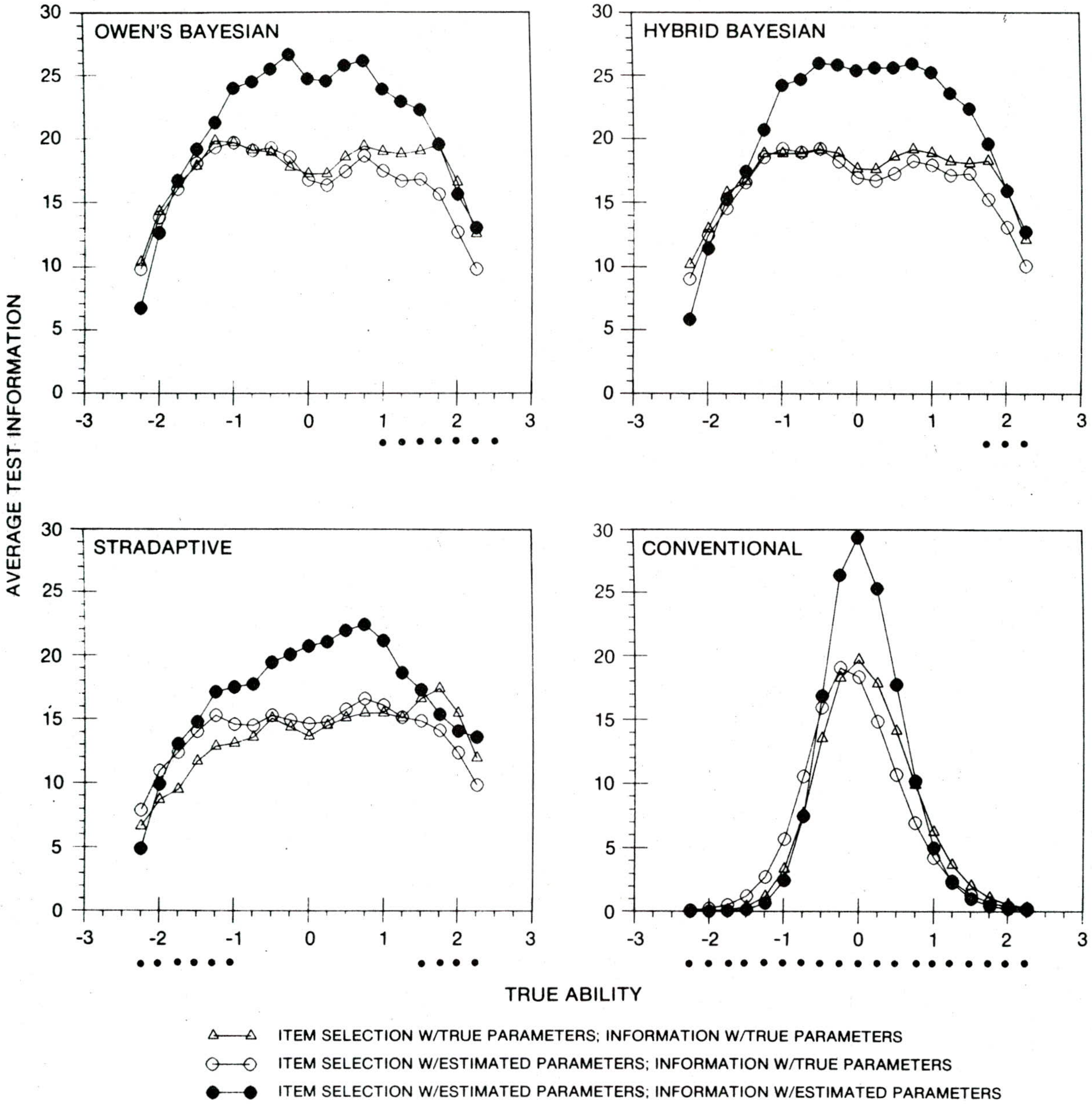


Figure 3. Best possible test information for a 15-item test, calculated from true and estimated item parameters.

Precision of Item Selection

Figure 4 compares item selection based on either true or estimated parameters for each of the four test strategies. Average test information was calculated in terms of true parameters for both the ideal and realistic item-selection conditions (open symbols). The Owen's Bayesian and hybrid Bayesian strategies yielded higher values of test information when item selection was based on true parameters rather than on estimated parameters. This effect was more pronounced for abilities above zero. The STRADAPTIVE test shows the same effect above an ability level of 1.0 but the reverse effect below that. The peaked conventional test shows a peaked information function at zero when item selection was based on true parameters, with rapidly declining information as ability diverges. When the peaked test item selection was based on estimated parameters, information based on true parameters is shifted to the left because of the systematic overestimation of the $\underline{b}$ parameter (i.e., it is peaked 0.25 lower than 0.0).

The line formed by the closed symbols in each panel of Figure 4 depicts the condition where item selection during testing was based upon estimated parameters, with test information being calculated with estimated instead of true parameters. This line is the information that would be indicated for the only parameters known in a practical "live" testing situation. This condition illustrates that the overestimation of item parameters is reflected in the estimated test information obtained for each simulation.



Note. The dots below the panels indicate significant differences between the two open symbol item selection conditions ($p < .05$ by Tukey HSD tests; Kirk, 1968).

Figure 4. Average test information obtained with four test strategies.

Figure 5 depicts the average test information for all four test strategies under ideal and realistic conditions. Although true and estimated parameter item-selection conditions are plotted separately, information for both panels was calculated with a common set of true parameters. Figure 5 shows that the difference between the two item-selection conditions was small compared to differences among the four test strategies. In both item-selection conditions, the two Bayesian tests were superior to the other tests over a wide range of ability. The hybrid Bayesian test information was only slightly less than that of Owen's Bayesian test, even though the latter required substantially more computation. The STRADAPTIVE test information increased with increasing ability, achieving the level of the other adaptive tests only for abilities above 2.0. Finally, the conventional test yielded high information near ability 0.0 but was clearly inferior to the adaptive tests elsewhere.

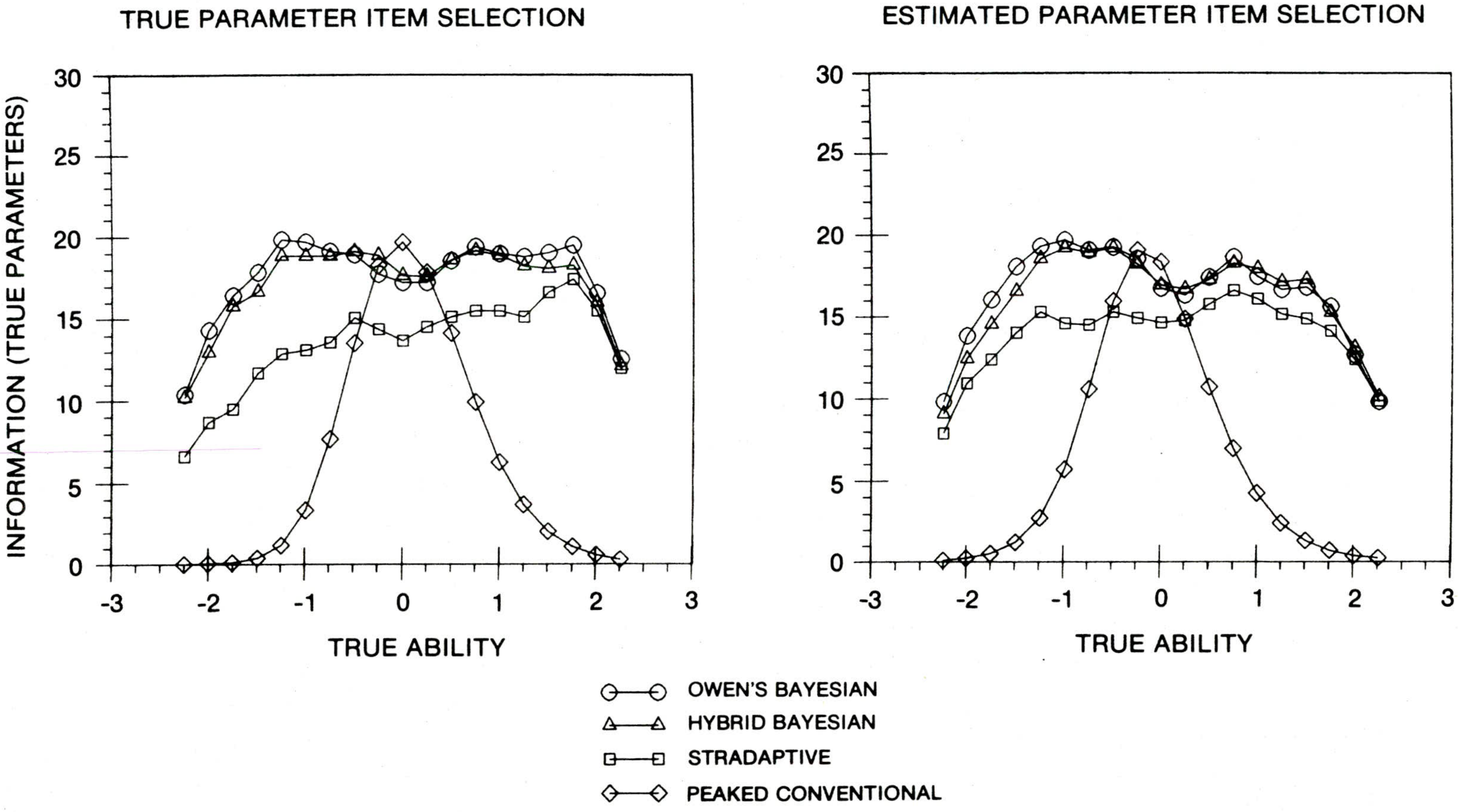


Figure 5. Average test information for each test using true vs. estimated item parameters for item selection.

Adaptive efficiency ratios were calculated to examine the proportion of the best possible test information achieved by each test strategy. The adaptive efficiency ratios plotted in Figure 6 are the result of dividing the obtained information values in Figure 5 by the corresponding best possible information with true parameters depicted in Figure 3 (lower curve). The Bayesian and hybrid tests were uniformly the most efficient with this measure, achieving about 80 percent of the best possible test information. The STRADAPTIVE test increased from about 50-60 percent efficiency at low abilities up to the level of the Bayesian tests at high abilities. The peaked conventional test efficiency was comparable to that of the adaptive tests only in a very narrow mid-range of the ability scale. For the ideal (true parameter) item selection condition, 100 percent efficiency was obtained at ability 0.0 because the conventional test was specifically constructed using the best possible items at that point.

Effective Test Length

Figure 7 shows the relative efficiency of the three adaptive test strategies under the realistic condition (estimated parameters), as compared with the ideal condition (true parameters). The left vertical axis is the ratio of the realistic condition test information to the ideal condition test information for each of the three adaptive strategies (i.e., dividing the lines in the right panel by those in the left panel of Figure 5). Ratios below 0.0 indicate proportionately less information in the realistic situation. The right vertical axis of the figure has been rescaled in terms of effective test length (i.e., multiplying the above ratio times the 15-item test length). The latter measure indicates that a testing strategy using estimated parameters for item selection was effectively a test of the length indicated, relative to the same strategy using true parameters. For the two Bayesian tests, the results indicate that approximately 1-3 items were needed for tests of above average examinees to achieve the test information that would be achieved if true parameters could be used for testing.

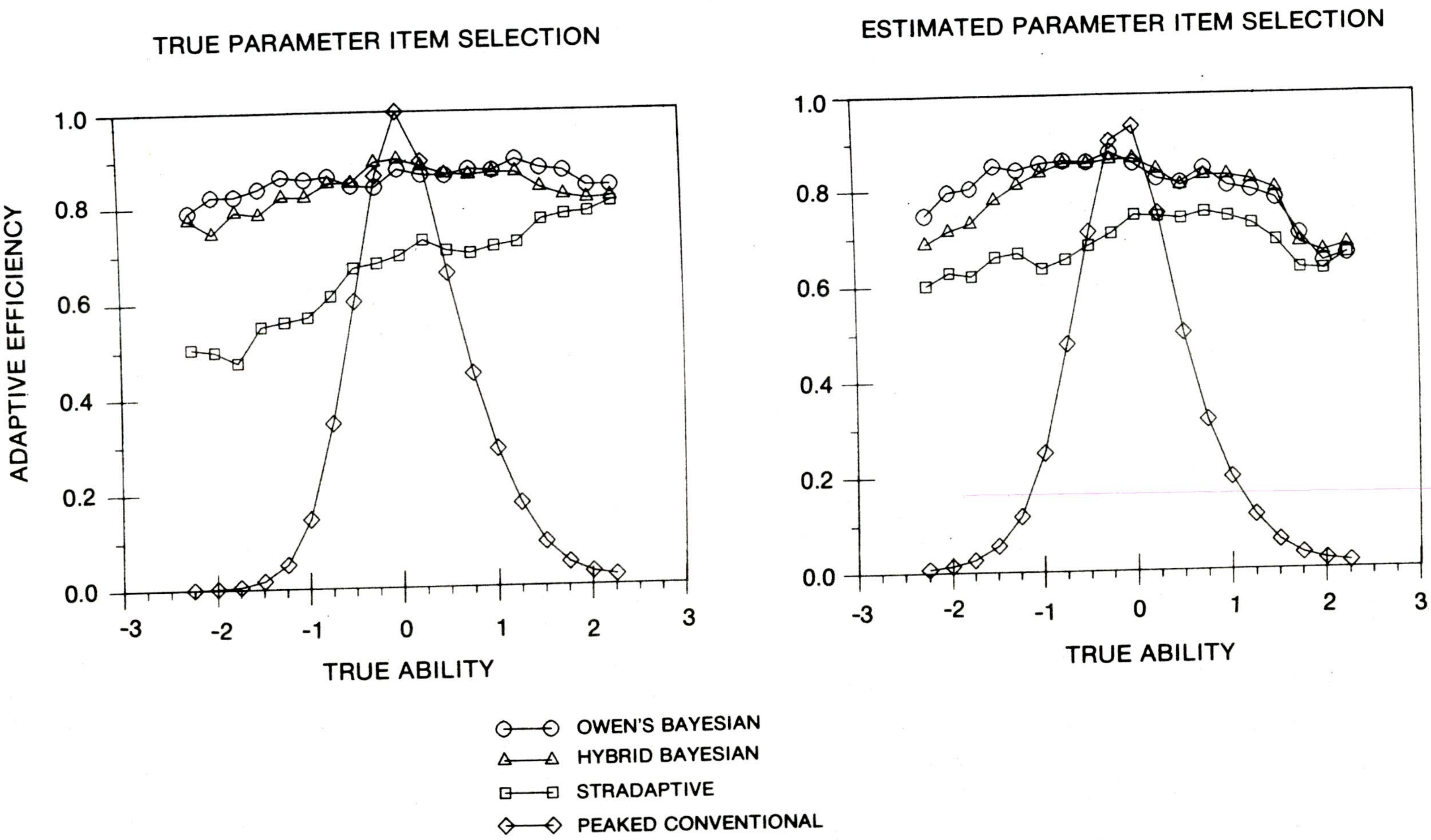


Figure 6. Adaptive efficiency for each test using true vs. estimated item parameters for item selection.

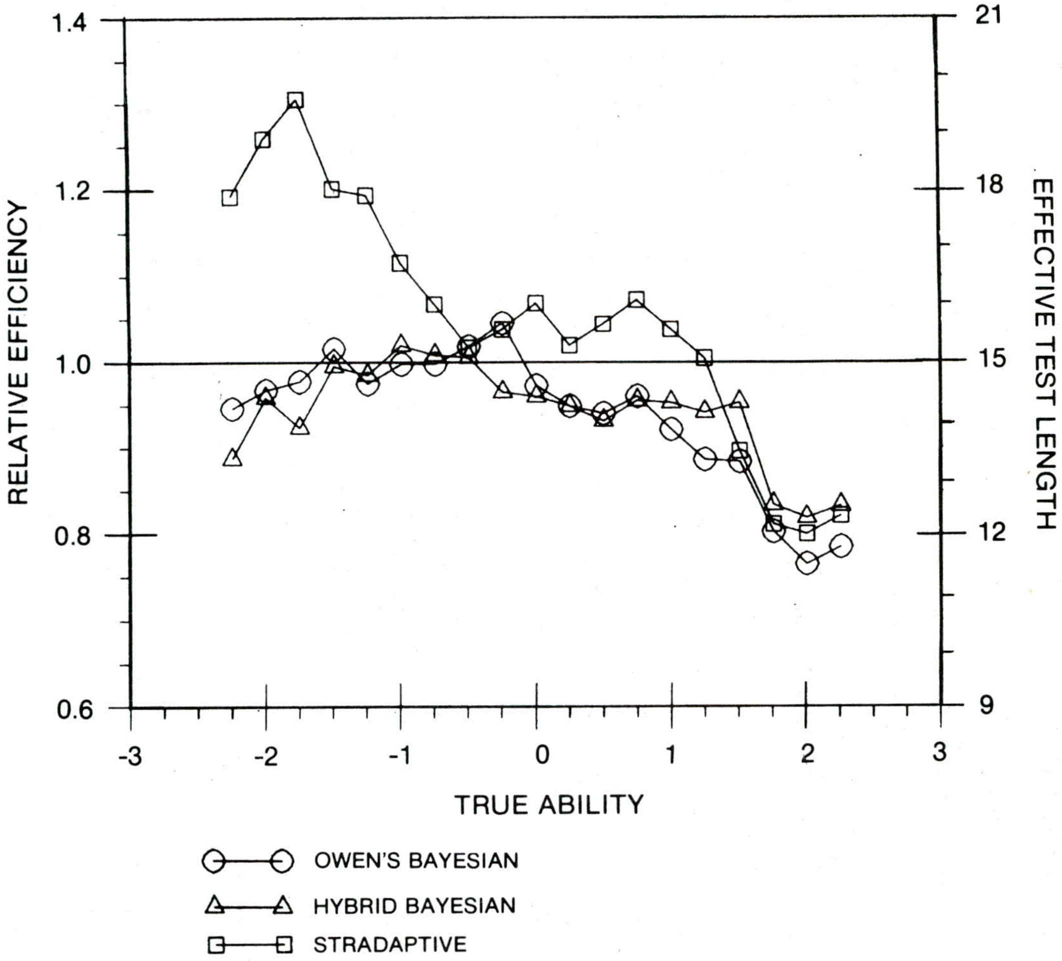


Figure 7. Relative efficiency and effective test length of three adaptive tests.

DISCUSSION

In the simulation study reported here, item parameters were overestimated by the estimation program OGIVIA. This resulted in test information being generally overestimated, when calculation of test information was based on estimated rather than true parameters. The role of fallible item parameters was then assessed by simulating tests in which item-selection during testing was based on either true or estimated parameters. These two conditions were compared in terms of average test information yield calculated from the known true parameters. This design allowed the effects of item-parameter estimation error to be assessed in three adaptive test strategies that varied in the extent to which item information was used to select test items.

The two Bayesian adaptive strategies made explicit use of item information during item selection. These two strategies were not disproportionately affected relative to tests making less optimal use of the information function for item selection. The two Bayesian tests also yielded the most measurement precision, in terms of either absolute level of test information or adaptive efficiency. Test information generally increased monotonically with ability in the case of the STRADAPTIVE test. While this test used an item pool that was prestratified on the basis of item difficulty and discrimination, the guessing parameter was not used. Its low test information at low abilities probably reflects the failure of the STRADAPTIVE test to account for guessing, since chance successes led to branching up to more difficult strata. Finally, the peaked conventional test yielded good precision only at the point where information had been concentrated prior to testing. Measurement precision declined rapidly outside this narrow band.

The present study focused on a few of many possible test strategies and also introduced a new test strategy. The hybrid Bayesian test was examined because it offered a means to achieve the good measurement characteristics of the full Owen's Bayesian procedure without the intensive computational requirements of the latter. It is worth noting that, since the hybrid test performed practically as well as the Owen strategy, it may therefore be considered as a possible candidate for implementation. Although not detailed here, the simulation study reported herein was also conducted for a maximum-likelihood information-table test (cf. Lord, 1977; Sympson, Weiss, & Ree, 1982) for stratified Bayesian tests (McBride, 1976b) and for a "flat" conventional test. Analyses of these other tests yielded similar conclusions about the effects of item-parameter estimation errors. Data similar to those presented here need to be developed using other item calibration programs, where the pattern of strong positive bias in all three parameters is less.

CONCLUSIONS

Item selection based on fallible item parameters resulted in little degradation of test information compared to item selection based on true item parameters. Adaptive testing strategies that explicitly optimize item information were not degraded more than test strategies making less optimal use of item information. It appears that errors in the item parameter estimates do not mitigate the psychometric advantages of these "optimal" strategies.

RECOMMENDATIONS

It appears that adaptive test strategies that explicitly make optimal use of item information offer more precise measurement than simpler strategies, even when item response model parameters are fallibly estimated. Therefore, rapid and secure methods that preserve the psychometric properties of these strategies should be developed and evaluated for use in the contemplated operational CAT system.

REFERENCES

Birnbaum, A. Some latent-trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Crichton, L. I. Effect of error in item parameter estimates on adaptive testing (Unpublished doctoral dissertation). Minneapolis: University of Minnesota, Psychology Department, Psychometric Methods Program, 1981.

Croll, P. R. Computerized adaptive testing system design: Preliminary design considerations (NPRDC Tech. Rep. 82-52). San Diego: Navy Personnel Research and Development Center, July 1982. (AD-A118 495)

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. Applied Psychological Measurement, 1977, 1, 111-120.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, CA: Brooks-Cole, 1968.

Lord, F. M. A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1977, 1, 95-100.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

McBride, J. R. Research on adaptive testing, 1973-1976: A review of the literature (Unpublished paper). Minneapolis: University of Minnesota, Psychology Department, Psychometric Methods Program, 1976. (a)

McBride, J. R. Simulation studies of adaptive mental tests: A comparative evaluation (Unpublished doctoral dissertation). Minneapolis: University of Minnesota, Psychology Department, Psychometric Methods Program, 1976. (b)

McBride, J. R. Computerized adaptive testing project: Objectives and requirements (NPRDC Tech. Note 82-22). San Diego: Navy Personnel Research and Development Center, July 1982. (AD-A118 447)

Owen, R. J. A Bayesian approach to tailored testing (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service, 1969.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Ree, M. J. Estimating item characteristic curves. Applied Psychological Measurement, 1979, 3, 371-385.

Schmidt, F. L., & Gugel, J. F. The Urry item-parameter estimation technique: How effective? (PS 76-1). Washington, DC.: U.S. Civil Service Commission, Personnel Research and Development Center, 1976.

Swaminathan, H., & Gifford, J. Estimation of parameters in the 3-parameter latent trait model. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, 1980.

Sympson, J. B., Weiss, D. J., & Ree, M. Predictive validity of conventional and adaptive tests in an Air Force training environment (AFHRL-TR-81-40). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, March 1982.

Urry, V. OGIVIA: Item parameter estimation program with normal ogive and logistic three-parameter model options. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, 1977.

Urry, V. W., & Dorans, N. J. Tailored testing, its theory and practice (NPRDC Tech. Rep.). San Diego: Navy Personnel Research and Development Center, in press.

Vale, C. D., & Weiss, D. J. A simulation study of STRADAPTIVE ability testing (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1975.

Waters, B. K. An empirical investigation of the STRADAPTIVE testing model for the measurement of human ability (Unpublished doctoral dissertation). Tallahassee: Florida State University, 1974.

Weiss, D. J. The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Program, September 1973. (AD 768376)

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1974.

DISTRIBUTION LIST

Assistant Secretary of Defense (Manpower, Reserve Affairs & Logistics)
Deputy Under Secretary of Defense for Research and Engineering (Research and Advanced Technology)
Military Assistant for Training and Personel Technology (ADUSD(R&AT))
Assistant Secretary of the Navy (Manpower & Reserve Affairs)
Principal Deputy Assistant Secretary of the Navy (Manpower and Reserve Affairs) (OASN(M&RA))
Deputy Assistant Secretary of the Navy (Manpower) (OASN(M&RA))
Director of Manpower Analysis (ODASN(M))
Chief of Naval Operations (OP-01), (OP-11), (OP-12) (2), (OP-13), (OP-14), (OP-15), (OP-115) (2), (OP-140F2), (OP-987H)
Chief of Naval Material (NMAT 0722), (NMAT 05)
Chief of Naval Research (Code 200), (Code 437), (Code 440) (3), (Code 442), (Code 442PT), (Code 458)
Chief of Information (OI-213)
Chief of Naval Education and Training
Commandant of the Marine Corps (MPI-20)
Commander of Chief U.S. Atlantic Fleet
Commander in Chief U.S. Pacific Fleet
Commander Naval Military Personnel Command (NMPC-013C)
Commanding Officer, Naval Aerospace Medical Institute (Library Code 12) (2)
Commanding Officer, Naval Regional Medical Center, Portsmouth, VA (ATTN: Medical Library)
Psychologist, ONR Branch Office
Commanding Officer, Office of Naval Research Branch Office, Chicago (Coordinator for Psychological Sciences)
Director Naval Civilian Personnel Command
Director, Naval Education and Training Program Development Center Detachment, Memphis
Superintendent, Naval Postgraduate School
Commander, Army Research Institute for the Behavioral and Social Sciences, Alexandria (PERI-ASL)
Headquarters Commandant, Military Enlistment Processing Command, Fort Sheridan
Headquarters, Air Force Military Personnel Center (Code MPCYPT)
Chief, Army Research Institute Field Unit, Fort Harrison
Commander, Air Force Human Resources Laboratory, Brooks Air Force Base (Scientific and Technical Information Office)
Commander, Air Force Human Resources Laboratory, Lowry Air Force Base (Technical Training Branch)
Commander, Air Force Human Resources Laboratory, Williams Air Force Base (AFHRL/OT)
Commander, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base (AFHRL/LR)
Superintendent, U.S. Coast Guard Academy
Defense Technical Information Center (DDA) (12)